# 線形化拡散写像手法の提案とその文書データへの適用

仲野 将† 立間 淳司† 青野 雅樹†

†豊橋技術科学大学情報工学系

## 1 はじめに

現在，世の中には様々なデータが存在している．その中でもテキストデータ，画像データなどは一般に非常に高次な特徴空間で表現される．これら高次元データに対して分類や検索等を行う場合，データをその高次元空間における特徴を保存した低次な空間で表現する次元削減を行うことで性能の向上が期待できる．

本研究では新しい線形次元削減の手法として線形化拡散写像を提案する．これは非線形な次元削減手法である拡散写像（DM:Diffusion Maps）[1]を新しい線形化アプローチであるスペクトル回帰 [2] により線形化したものである．

評価実験として英語の文書データである 20Newsgroups Dataset のサブセットに対して提案手法を適用し，Nearest Neighbor 及び DCG により評価する．またこれまでに提案された文書データに適用されている手法との比較を行い，提案手法の有効性を確認する．

## 2 関連研究

文書データに対する次元削減手法として有名なものに LSI（Latent Semantic Indexing）が挙げられる．LSI はベクトル空間モデルで表現された文書データに対して，特異値分解により特徴的な低次元表現を求める．近年，全体的には非線形構造を成していても，局所的にはユークリッド空間と同じであるという多様体の性質を利用した非線形次元削減手法が提案されている．LE（Laplacian Eigenmaps）はそれら手法の内の一つである．LPI（Locality Preserving Indexing）は LE を線形化して文書データに適用したものであり，LSI よりも良い性能が得られるとされている [3]．RLPI（Regularized LPI）は LE をスペクトル回帰により線形化して文書データに適用したものであり，LPI よりも良い性能が得られている [4]．

DM は多様体の性質を利用した非線形次元削減手法の一つであり，LE をより一般的にしたものと考えられる．我々は DM の線形化手法である LDP を提案している [5]．スペクトル回帰はこれまで行われてきている線形化アプローチとは違い，非線形手法により得た低次元表現に対して正則化項を付加した線形回帰を行うことにより射影行列を得る．DM をスペクトル回帰により線形化した研究は，我々が調べた限りまだ行われてはいない．

**Linearized Diffusion Maps and its Application to Documents**

Masaru NAKANO†, Atsushi TATSUMA†and Masaki AONO†
†Dept. of Information and Computer Sciences, Toyohashi University of Technology
441-8580, Toyohashi, Japan

## 3 提案手法

$m$ 次元 $N$ 個のデータ $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^m$ が与えられている．本手法により，$n \ll m$ である射影行列 $F \in \mathbb{R}^{m \times n}$ を求める．$X$ の低次元表現 $Y = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N] \in \mathbb{R}^n$ は $Y = F^T X$ として以下のステップで求める．

1. $X$ より重み付きグラフ $G = (V, E, W)$ を作る．ここでは文書データへの適用に際し，全ての頂点が到達可能な $k$ 近傍グラフを用いる．

2. グラフの重み行列 $W$ を求める．ここでは文書データへの適用に際し $W$ の要素 $w_{ij}$ はコサイン類似度を用いて次式より与えられる．

$$w_{ij} = cossim(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} \quad (1)$$

3. 異方性遷移カーネル行列 $P^{(\alpha)}$ を求める．

$$W^{(\alpha)} = Q^{-\alpha} W Q^{-\alpha}, \quad q_{ii} = \sum_j w_{ij}$$
$$P^{(\alpha)} = D^{-1} W^{(\alpha)}, \quad d_{ii} = \sum_j w_{ij}^{(\alpha)}$$

$Q$ 及び $D$ はそれぞれ $q_{ii}, d_{ii}$ を対角要素とする対角行列である．ここではデータのサンプリング密度による影響を抑えるため，$\alpha = 1$ とする．

4. 固有値問題 $P^{(\alpha)}\psi = \lambda\psi$ を解き，固有値 $1 = \lambda_0 > \lambda_1 \geq \ldots \geq \lambda_n$ を得る．非線形手法による低次元表現 $A$ の要素 $\mathbf{a}_i$ は $\lambda_j\,(j = 1, 2, \ldots, n)$ に対応する固有ベクトル $\psi_j\,(j = 1, 2, \ldots, n)$ を用いて $\mathbf{a}_i = \left[\lambda_1^t \psi_1^{(i)}, \lambda_2^t \psi_2^{(i)}, \ldots, \lambda_n^t \psi_n^{(i)}\right]^T$ となる．ここでは $t = 1$ とする．

5. L2 ノルム正則化項を用いた線形回帰により，射影行列 $F$ を求める．

$$F = \begin{cases} \left(XX^T + \beta I\right)^{-1} XA^T & if\ N > m \\ X\left(XX^T + \beta I\right)^{-1} A^T & otherwise \end{cases}$$

$\beta$ は回帰における正則化パラメータである．

6. 線形化拡散写像による低次元表現 $Y = F^T X$ を得る．学習時に存在しない新規データに対しても射影行列 $F$ を用いることにより低次元表現を求めることができる．

## 4 評価実験

提案手法の有効性を確認するために，評価実験を行う．データとして 20Newsgroups Dataset[1] 18828 バージョンを使用する．その中に含まれる 10 クラスより，各々500 個のデータをランダムに取り出し，合計 5000 個を使用する．これに対して統計テキスト処理ツール Rainbow[2]を用いて文書データよりベクトル空間モデルを作る．本実験ではこれにより 15246 次元の特徴ベクトルを得た．最後に tf-idf による重み付けを行い，各ベクトルの L2 ノルムが 1 となるように正規化を行うことにより，本実験で使用する実験データを得る．

実験では，まず実験データを各クラス 300 個ずつ計 3000 個の訓練データと，残りの各クラスから 200 個ずつ計 2000 個のテストデータにランダムに分割する．本実験ではランダムな分割による実験結果の差を抑えるために，10 個の訓練とテストの分割パターンを用意する．次に訓練データに対して，提案手法及び比較手法の各種線形次元削減手法を適用し，射影行列 $F$ を得る．このとき各手法で用いるパラメータは，訓練データによる交差検定を行い，評価基準において最も良い結果となった値を使用する．本実験では処理時間等の関係より，5 交差検定により，テストに使用するパラメータを求める．次に射影行列 $F$ を用いて訓練データ及びテストデータの低次元表現を求め，データ間の類似度により各評価基準でテストを行い比較する．データ間の類似度には式 (1) に示すコサイン類似度を用いる．評価基準には，データの分類精度を比較するための Nearest Neighbor 及び情報検索の精度を比較するための DCG（Discounted Cumulative Gain）を用いる．比較実験として文書データに用いられる手法である，LSI，LPI，及び RLPI との比較を行う．

Nearest Neighbor による実験結果を表 1 に示す．これは 10 パターンの分割による実験でそれぞれ得られた最大精度及びその時の次元数を平均したものである．括弧内の数字はそれぞれの標準偏差を表している．これより，次元削減によって文書分類の精度が向上し，その中でも提案手法が十分に低い次元数で最も高い精度を得ることができたということがわかる．

DCG による実験結果を表 2 に示す．これはそれぞれの分割において，順位 20 番目までの DCG が最大となるもの及びその時の次元数の平均を表している．括弧内の数字は標準偏差を表している．これより，次元削減により文書検索の精度が向上し，その中でも提案手法は全ての順位において最もよい DCG スコアを得ることができたということがわかる．

表 1: Nearest Neighbor による実験結果

| | Precision (%) | Dimension |
|---|---|---|
| Baseline | 79.74 (±0.7) | 15246 (−) |
| LSI | 81.26 (±0.7) | 806 (±137) |
| LPI | 82.72 (±0.5) | 45 (±14) |
| RLPI | 83.45 (±0.4) | 58 (±12) |
| Proposal | **84.90** (±0.4) | 77 (±19) |

表 2: DCG による実験結果（1, 10, 20）

| | $DCG_1$ | $DCG_{10}$ | $DCG_{20}$ | Dimension |
|---|---|---|---|---|
| Baseline | 0.7974 | 3.486 | 4.765 | 15246 (−) |
| LSI | 0.7689 | 3.742 | 5.380 | 29 (±3.0) |
| LPI | 0.8142 | 4.047 | 5.884 | 29 (±5.4) |
| RLPI | 0.8244 | 4.125 | 6.012 | 26 (±6.6) |
| Proposal | **0.8346** | **4.188** | **6.117** | 33 (±4.6) |

## 5 おわりに

本研究では非線形次元削減手法である DM を正則化項を付加した線形回帰により線形化を行うスペクトラル回帰により線形化した新しい次元削減手法を提案した．評価実験では DM の持つデータのサンプリング密度への依存を抑えるという性質等の特徴により，良い結果を得ることができたと考えられる．

今後の課題として，教師なし学習である拡散写像の持つ良い性質を生かしつつ，教師あり，半教師あり学習へ拡張する方法の開発などが考えられる．また他のデータセットに対しても同様の実験を行い，提案手法の有効性を確認することが必要である．

## 参考文献

[1] R. R. Coifman *et al.*: "Diffusion maps", *Applied and Computational Harmonic Analysis*, Vol. 21, No. 1, pp. 5-30 (Jul. 2006).

[2] D. Cai *et al.*: "Spectral Regression for Dimensionality Reduction", *Dept. of Computer Science, UIUC Tech. Report*, No. 2856 (May 2007).

[3] X. He *et al.*: "Locality preserving indexing for document representation", *ACM SIGIR* , pp. 96-103 (Jul. 2004).

[4] D. Cai *et al.*: "Regularized Locality Preserving Indexing via Spectral Regression", *ACM CIKM*, pp. 741-750 (Nov. 2007).

[5] 立間淳司, 仲野将, 青野雅樹: "線形拡散射影と三次元物体の形状類似検索への応用", 論文投稿中.

[1] http://people.csail.mit.edu/jrennie/20Newsgroups/
[2] http://www.cs.umass.edu/~mccallum/bow/rainbow/